IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

Application of Big Data Tools and Techniques in Prediction of Heart Diseases

R.Sharmila[1*], S.Chellammal[2]

[1,2]Bharathidasan University Constituent Arts & Science College, Navalurkuttpattu, Trichy – 620027, TamilNadu, India

[*]*Corresponding Author: sharmiparam@gmail.com;chelsganesh@gmail.com*

Available online at: www.ijcseonline.org

Abstract— Heart disease is one of the major causes of death in human life. But, prediction of heart disease with desired accuracy is difficult due to many reasons. For example, the database of heart disease is being archived for many years with huge volume which are too large for traditional systems to process. Also, the clinical reports and medical tests related to heart disease produce in a variety of formats such as text, images, sound etc which are not effectively handled by traditional database systems. Nowadays data mining algorithms and big data technologies play crucial role in the prediction of heart diseases. In addition, big data techniques are useful in finding the patterns of heart disease in its early stage. Analysis of heart disease data can be done on big scale using Hadoop, R and MapReduce. In our previous paper, we proposed a conceptual approach for prediction of heart diseases with Support Vector Machine (SVM) in parallel programming fashion. In relation to our previous work, in this paper, an investigation is done in finding the applicability of different big data tools and technologies for prediction of heart diseases.

Keywords: Big data tools, Hadoop, Map Reduce, prediction of diseases

I. INTRODUCTION

Prediction of heart disease is one of the major challenges in health care industry [1]. Traditionally diseases have been predicted manually [2]. In order to facilitate the prediction of diseases, data mining algorithms have been in use for the past couple of decade [3]. Though mining techniques provide significant contribution in providing higher accuracy [4], such techniques have to be combined with recent techniques and technologies in order to meet the evolving data growth and its nature. For example, in conventional heart predicting systems, data mining algorithms have used with heart disease data which was highly structured (i.e. most of the data occur in rows and columns). But nowadays the medical equipments and tests produce data in different formats and they are unstructured. For example, the output of angiogram, angioplasty, Cardiac Computed Tomography, etc are images in one dimension to three dimensions. Though conventional technologies such as relational database managements are yielding higher accuracy with standard data, they are not able to handle unstructured data. Further, conventional systems can handle data of size upto 1TB[5]. But the analysis of data more than 1TB becomes more common requirement in healthcare industry. Another factor to be considered is that the speed with which data arrives. For example, in the case of telemedicine [6], different parameters are sensed at regular intervals and sent to a remote location for further diagnosis. In this example, data occurs with some speed. In Intensive Care Unit or Critical Care Unit, data is getting generated with high speed. Hence, it becomes essential to process those data instantly so that the proper treatment can be given.

In nutshell, compared to traditional method of analyzing and predicting, data, it is becomes an urgent need to analyze the existing data with big data and techniques. Further, in this aspect, in [7], an approach is proposed to enhance prediction of heart diseases using data mining techniques in big data environment. This method uses Support Vector Machine (SVM) in parallel fashion in a distributed storage system. It is proposed to store the data in distributed nodes of Hadoop Distributed File System (HDFS). It is proposed to employ MapReduce programming model for simultaneously processing. In this perspective, it is proposed to study role and applicability big data tools and technologies in prediction of heart diseases. In this paper, the need for big data tools and techniques is described in the introduction. Section II describes existing medical tests for heart disease, their output data formats and typical size of data. In Section III some tools are identified according to the data formats mentioned in Section II. Section IV concludes the paper.

II.NATURE OF EXISTING DATA FORMATS

Different kinds of medical tests are used to detect the presence of heart disease in a person. Typical tests, their data formats and approximate size of data involved are given in Table 1.

Table 1 Different kinds of tests, output data formats and their typical size

S.No		Output Data Format	Size of Output
1.	Blood test	Text	9.23MB
2.	Holter Monitor	Text + Table +graph	10.7KB
3.	Stress test	Table form	361KB
4	Echo	Text + Figure	254KB
5	ECG	Graph	2.77MB
6	CT scan	Image	18.4KB
7	Cardiac Cathetherization	Text + Table +Figure	914KB
8	Heart disease information in the form of tweets/SMS	Text	140 characters/bytes
9	Cardiac MRI	Image	72.5KB
10	Angiogram	Image	128KB
11	ICU/CCU data	Graph	Not archived
12	Telemedicine sensor data	Numeric/text/images	Of the order of KB
13	Standard attributes for prediction (age,sex,cp,trestbps,chol,restecg,fbs,thalach, Exang,oldpeak,slope,ca,thal,num)	Numeric	34.7MB
14	Electronic Health Record	Text	Of the order of KB

From Table 1, it is understood that along with static data, data is generated at regular intervals also. For example, in the case telemedicine, typical parameters such as heart rate, blood pressure, temperature, other vital signs, images and videos are being monitors and sent in real to Doctor's site. This helps the physician for any treatment according the conditions of the patient. So, in this case, the data is being acquired regularly. In a similar manner, in ICU and CCU, critical parameters are being monitored continuously.

From Table 1, it is understood that along with static data, data is generated at regular intervals also. For example, in the case telemedicine, typical parameters such as heart rate, blood pressure, temperature, other vital signs, images and videos are being monitors and sent in real to Doctor's site. This helps the physician for any treatment according the conditions of the patient. So, in this case, the data is being acquired regularly. In a similar manner, in ICU and CCU, critical parameters are being monitored continuously.

III.BIG DATA TOOLS FOR PREDICTION OF HEART DISEASES

It is found that the data involved in predicting heart diseases are of image, videos, text, standard data in rows & columns as in the case of EHR, etc. In addition, the data occurs with some velocity as in the case of telemedicine, ICU and CCU. Regarding archival, most of the images are not archived in hospital. Reports in the form of images are given to concerned patients and they need to carry their reports further visits to hospitals. If it is essential to store all the images for medical follow-up tasks, huge amount of data also will need to be stored. Another aspect to be considered is that whether the parameters observed during ICU or CCU is to be archived for later analysis. In general, ICU data or CCU data are not archived. But archival of such data is essential for better diagnosis even of the same patient or similar kinds of patients. The possibility of archival ICU/CCU data needs to be explored. Probably not all the ICU/CCU parameters need not to be archived. In case of abnormalities, it becomes essential to store such data. This requires tools which facilitate real time analytics. Whenever some data arrives, analysis has to be made to decide whether the data is to be stored or not. According to the decision, data will be archived for later analysis or not archived. From the data, their format and size some tools are identified as given in Table 2.

Table 2.

Data	storage	Processing techniques
Text	Hive	Text analytics tools
Image	disk /NoSQL	Image processing tools like Matlab
Rows/Columns (for example EHR)	Hbase/Cassandra	Distributed
ICU/CCU	HDFS	Apache Spark for both batch as well as streaming data processing
Large data	HDFS	Parallel programming paradigm using tools such as MapReduce

The applicability of different big data tools for heart disease prediction is as shown in Fig. 1. A small overview of these tools is as follows. To store large data, HDFS is predominantly used. HDFS stores data in different nodes, called data nodes. Name node is responsible for splitting and distributing the data into different data nodes. Once the data is distributed into different data nodes, it is essential to bring processing of data distributed into different nodes. Parallel programming can be brought using MapReduce.

HDFS and MapReduce are bundled together in Hadoop. Job tracker assigns the required processing task in parallel fashion with MapReduce to task tracker. The combination of HDFS and Mapreduce facilitate the offline or batch processing of huge data. Whenever there is a need for real data analysis of says for example ICU data, the Spark tool can be used. It is much faster than MapReduce. In addition, Spark has different packages such as spark streaming, spark sql, spark mllib to facilitate different kinds of tasks. Spark can handle data arrives at speed. Further, there are tools such as Cassandra and MongoDB which are basically schemaless databases and they can be used to store any kind of data such as image. Typically images of ECG, MRI can be stored and the knowledge of those images can be obtained from image processing tools such as Matlab. Thus, rich knowledge can be obtained not only from big data tool but from other disciplines as shown in Fig. 2. Big data tools are used along with data mining algorithms, image processing tools, text analytics tools, statistics, data science, etc.

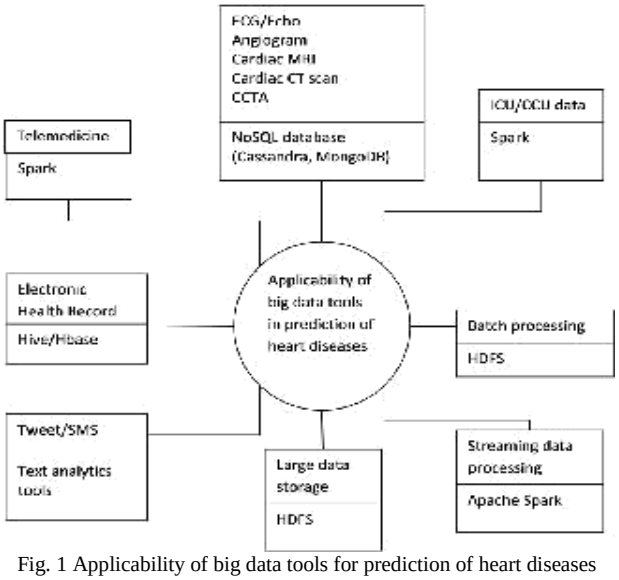

Fig. 1 Applicability of big data tools for prediction of heart diseases

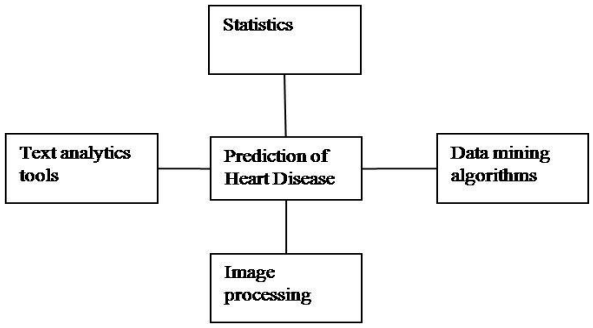

Fig. 2 Big data tools with other disciplines for heart disease prediction

IV.CONCLUSION

Prediction of heart disease is one of the challenging issues in healthcare industry. In this paper, the applicability of different big data tools for the prediction of heart disease is studied. It is understood that the combination of big data tools with other existing techniques such as text analytics tools, data mining, statistics and image processing will certainly provide higher prediction accuracy.

REFERENCES

[1]. Revathi.T, Jeevitha*,"Comparative Study on Heart Disease Prediction System Using Data Mining Techniques",*International Journal of Scienc and Research,Volume:4,Issue 7,July 2015,PP.2120-2123.

[2]. Miss.Chaitrali S.Dangare, Dr.Mrs.Sulabha S.Apte*,"A Data Mining Approach for Prediction of Heart Disease Using Neural Networks",*International Journal of Computer Engineering & Technology,Volume:3,Issue 3,December 2012,pp.30-40

[3]. Thenmozhi.K, Deepika.P," *Heart Disease Prediction Using Classification with Different Decision Tree Techniques",* International Journal of Engineering Research and General Science,Volume:2,Issue 6,November 2014,pp.6-11.

[4]. Subha.V, Revathi.M, Murugan.D,"Comparative *Analysis of Support Vector Machine Ensembles for HeartDiseasePrediction*",InternatinalJournalofComputerScience& CommunicationNetworks,Volume:5(6),December 2015,pp.386-390

[5]. Deepali Aggarwal, "Differences between traditional data and big data",June30,2016

[6]. Andrea Di Lenarda et al, *"The future of telemedicine for the management of heart failure patients* ",Journal of the European Society of Cardiology, May 2017

[7]. Sharmila.R, Chellammal.S*,"A conceptual method to enhance the prediction of heart disease using big data techniques",* International Journal of Computer Science and Engineering,Volume:6,Issue 4,May 2018,pp.25.

